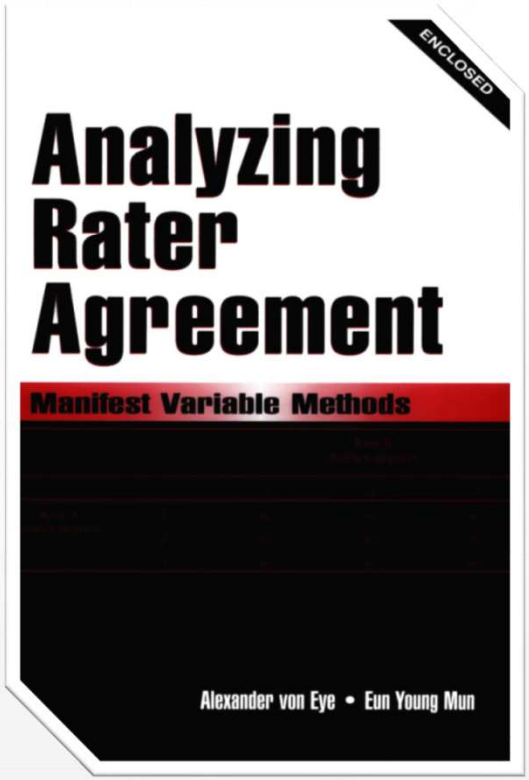
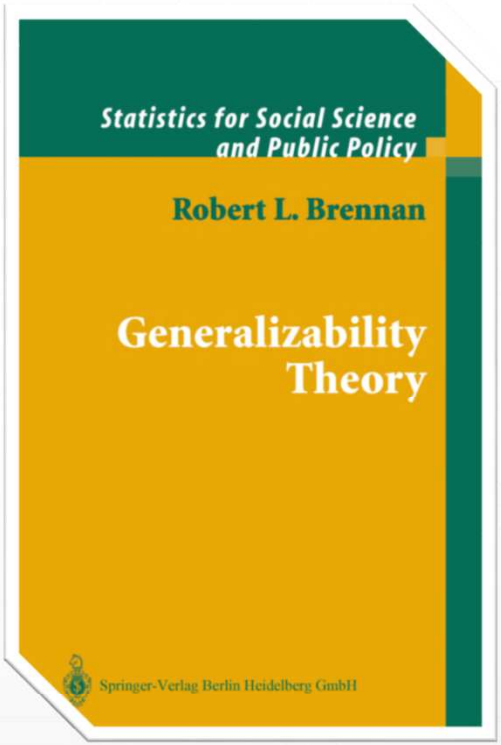
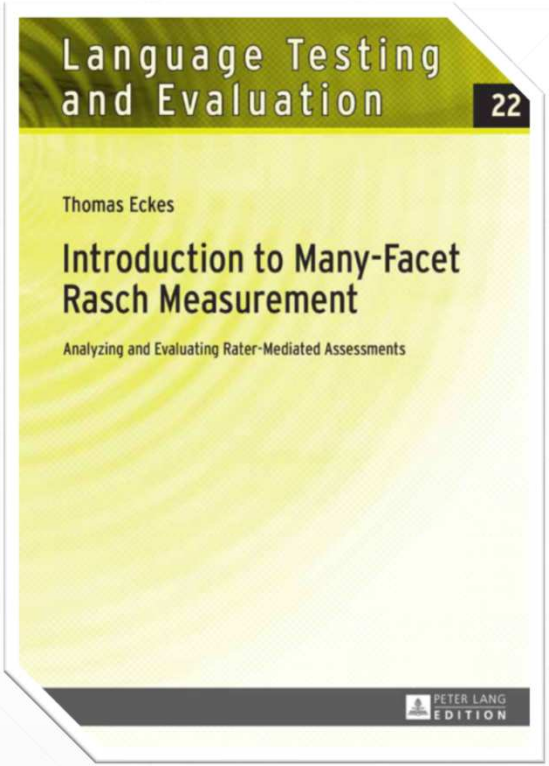
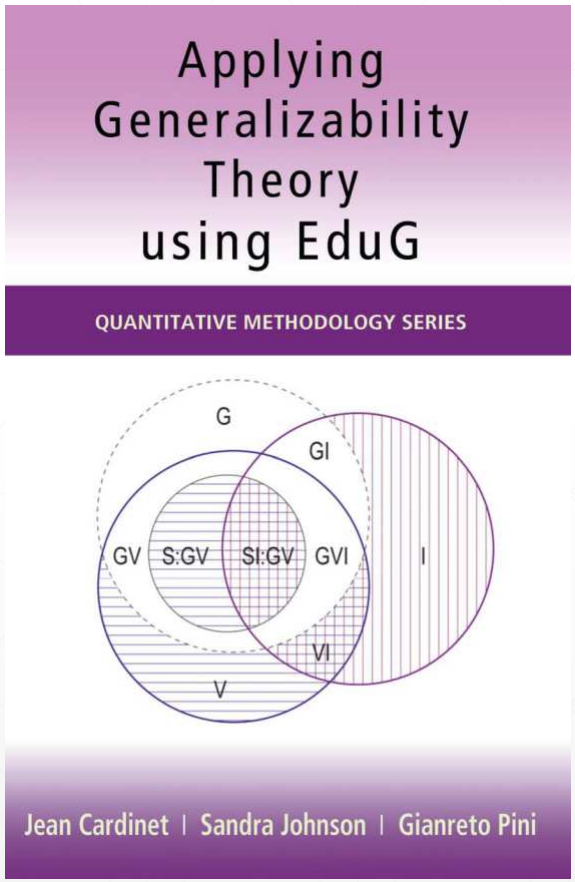
Melibatkan Rater dalam Pengukuran

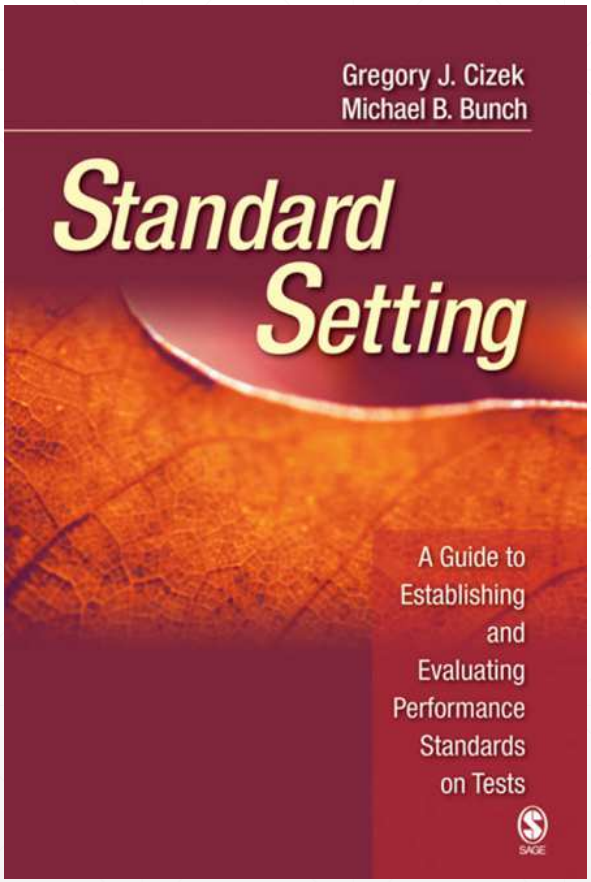
Wahyu Widhiarso | Fakultas Psikologi UGM


ENCLOSED
Analyzing Rater Agreement
Manifest Variable Methods
Alexander von Eye • Eun Young Mun


Statistics for Social Science and Public Policy
Robert L. Brennan
Generalizability Theory
Springer-Verlag Berlin Heidelberg GmbH


Language Testing and Evaluation
22
Thomas Eckes
Introduction to Many-Facet Rasch Measurement
Analyzing and Evaluating Rater-Mediated Assessments
PETER LANG EDITION


Applying Generalizability Theory using EduG
QUANTITATIVE METHODOLOGY SERIES
G
GI
GV
S:GV
SI:GV
GVI
I
VI
V
Jean Cardinet | Sandra Johnson | Gianreto Pini


Gregory J. Cizek
Michael B. Bunch
Standard Setting
A Guide to Establishing and Evaluating Performance Standards on Tests
SAGE

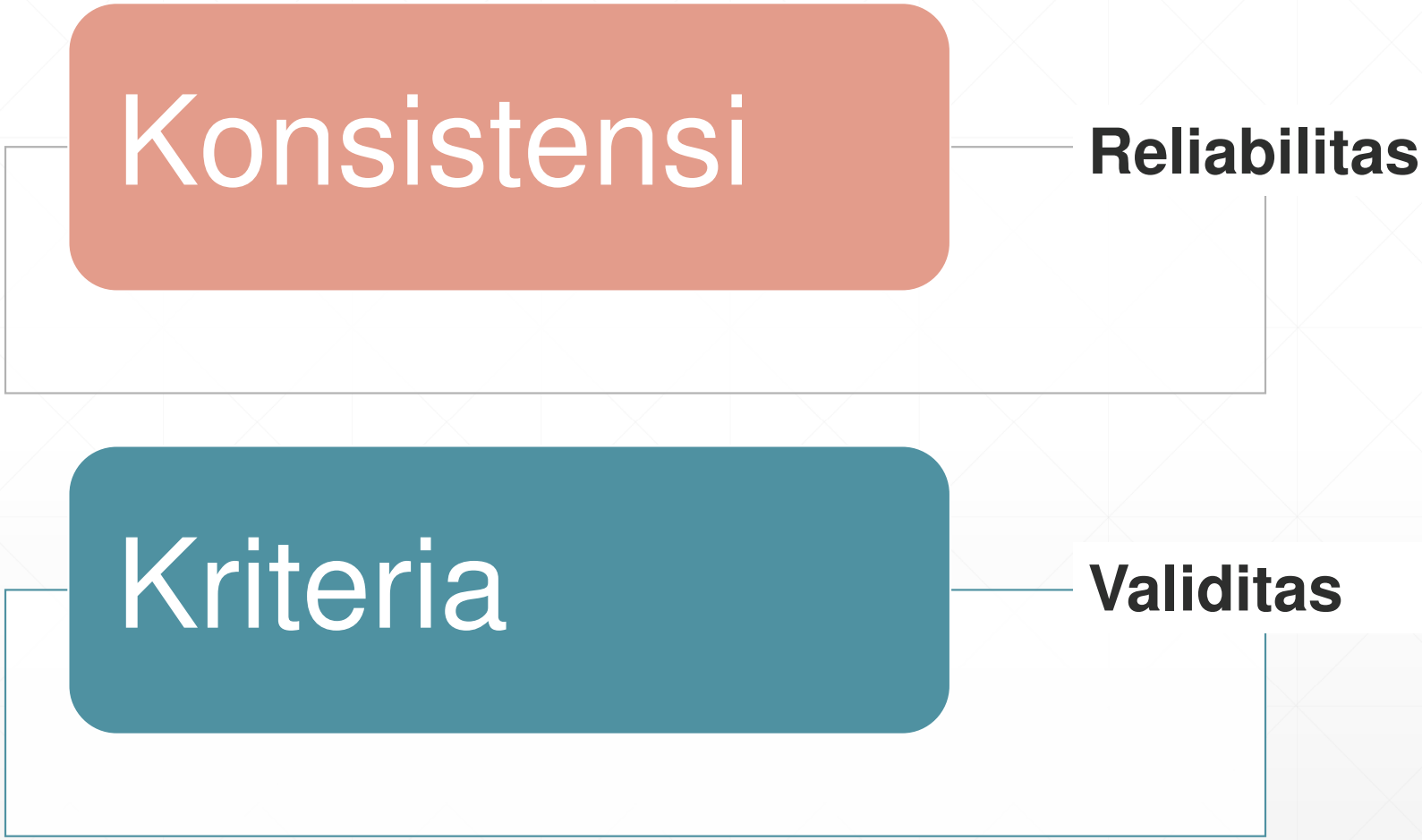
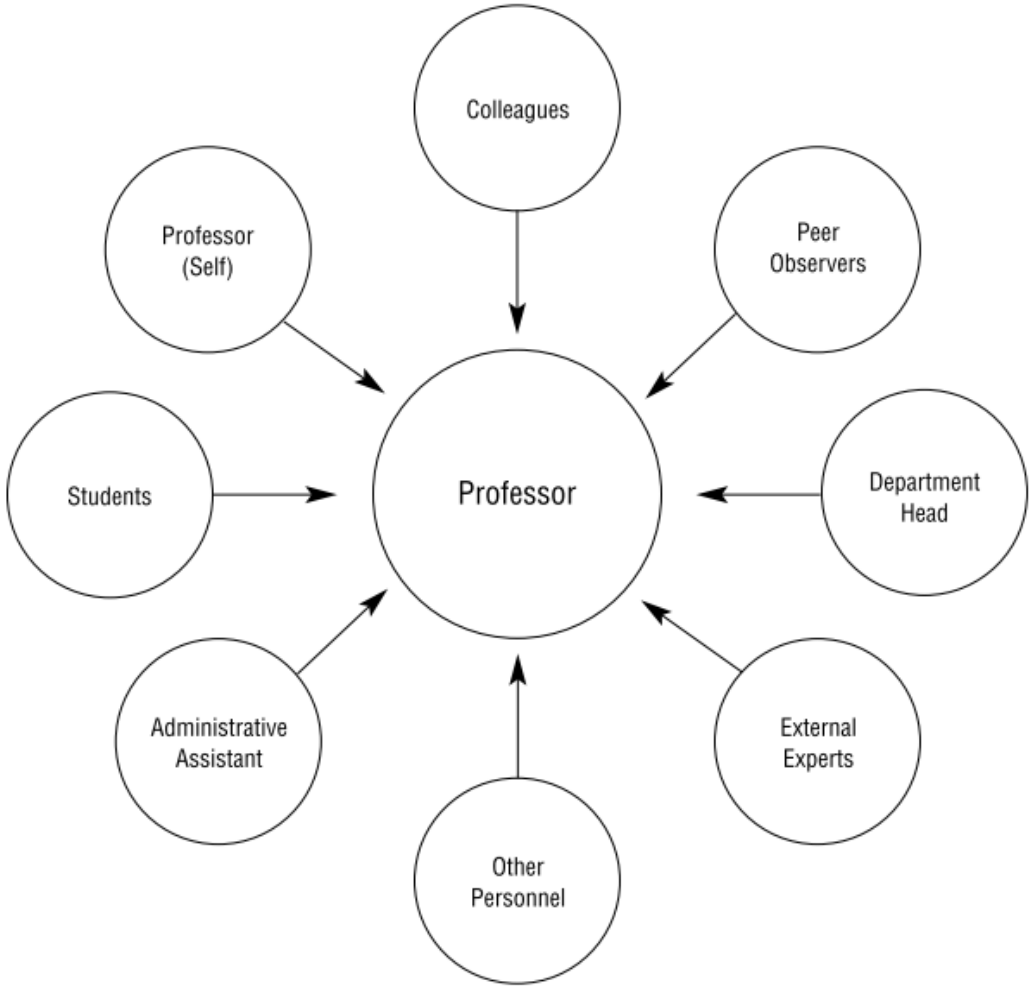

Konsistensi
Reliabilitas
Kriteria
Validitas

Figure 1.3 360° Multisource assessment of a professor.



Source: Me and You (2006b).

Kasus #1
Analisis Varians

Summary ANOVA

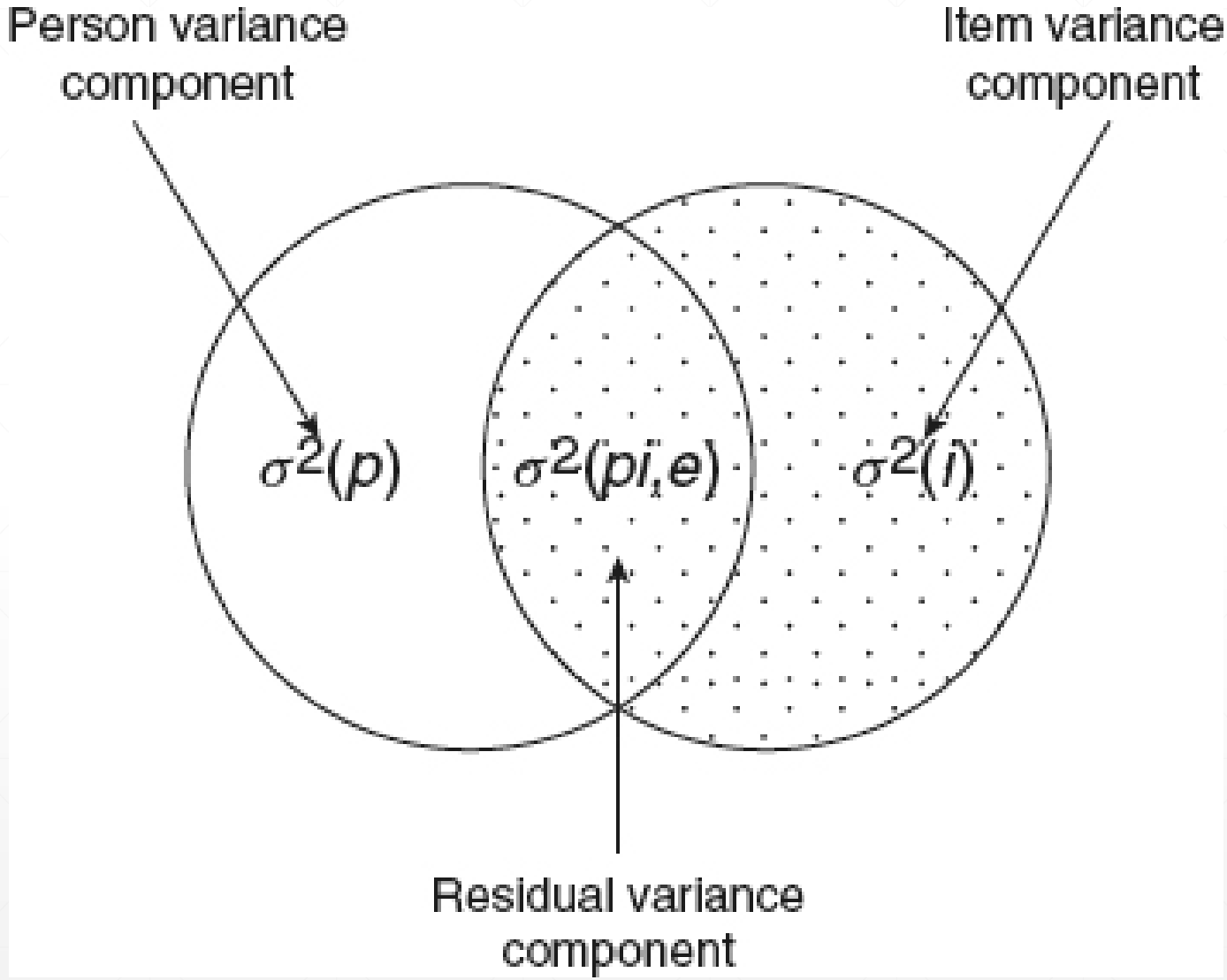
Source	Sum of Squares	Degrees of Freedom	Variance Estimate (Mean Square)	*F* Ratio
Between	SS_B	$K-1$	$MS_B = \frac{SS_B}{K-1}$	$\frac{MS_B}{MS_W}$
Within	SS_W	$N-K$	$MS_W = \frac{SS_W}{N-K}$	
Total	$SS_T = SS_B + SS_W$	$N-1$		

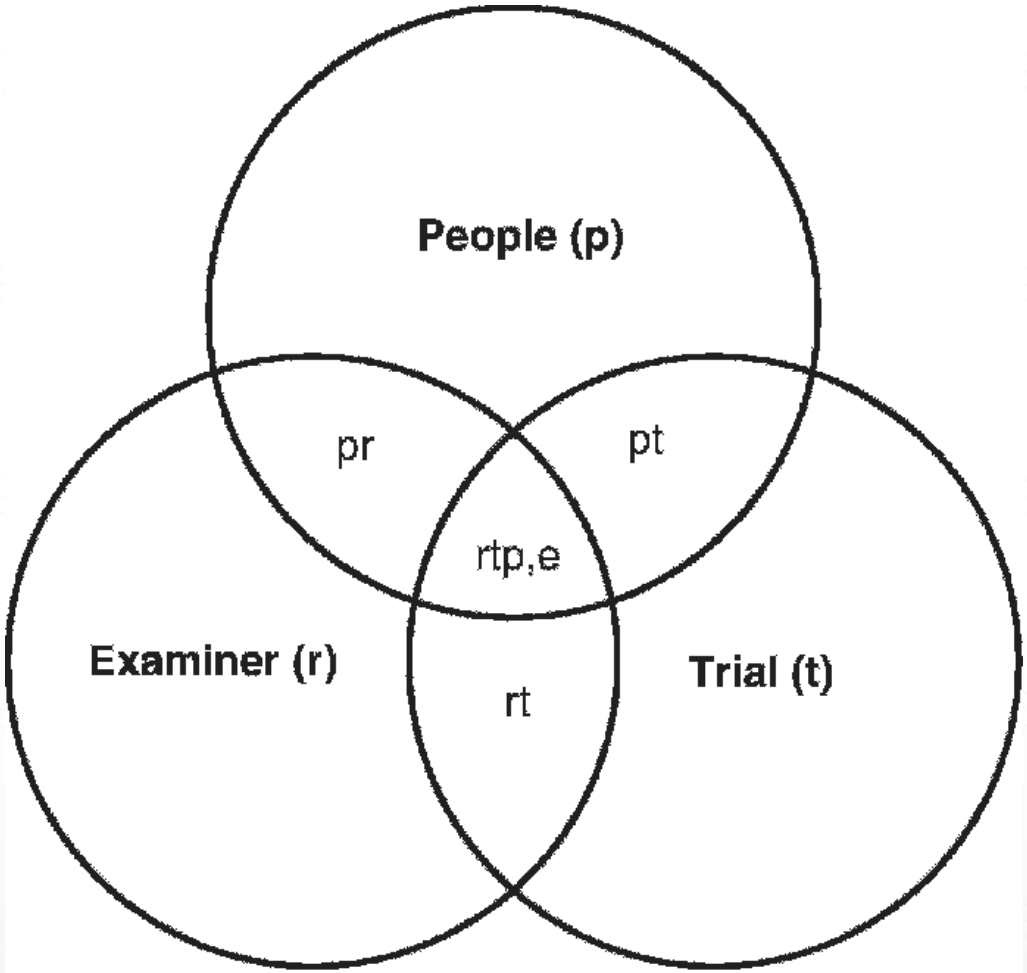
Source of variation	df	*F*
Dependent variable: earlywood tracheid length		
Treatment (A)	1	1.01
Distance (B)	1	161.69***
A × B	1	7.21**
Covariate	1	1.18
Dependent variable: lignin concentration		
Treatment (A)	1	0.03
Distance (B)	1	0.12
A × B	1	0.83
Covariate	1	0.82

Kasus #2
Generalizability Theory

	Postpartum hemorrhage	Preeclampsia
Variance explained (%)		
Simulation participating team	39.8	29.4
Rater	5.3	10.6
Item	0.9	0.1
Team*Rater	31.7	37.4
Team*Item	0.3	1.5
Rater*Item	1.5	1.9
Team*Rater*Item	20.5	18.1
Generalizability coefficient	0.83	0.76

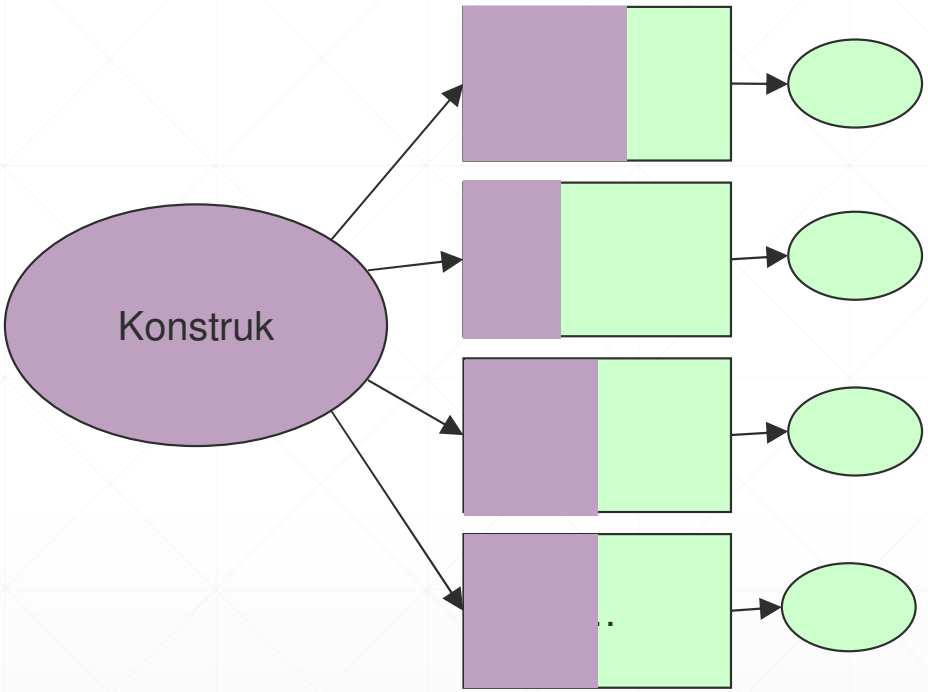
*Represents interaction.

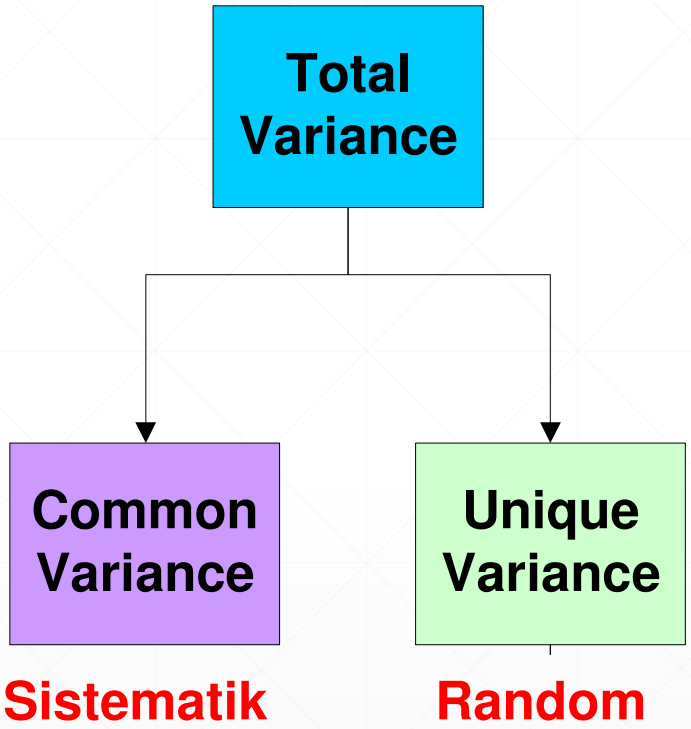

Person variance component
Item variance component
$\sigma^2(p)$
$\sigma^2(pi,e)$
$\sigma^2(i)$
Residual variance component



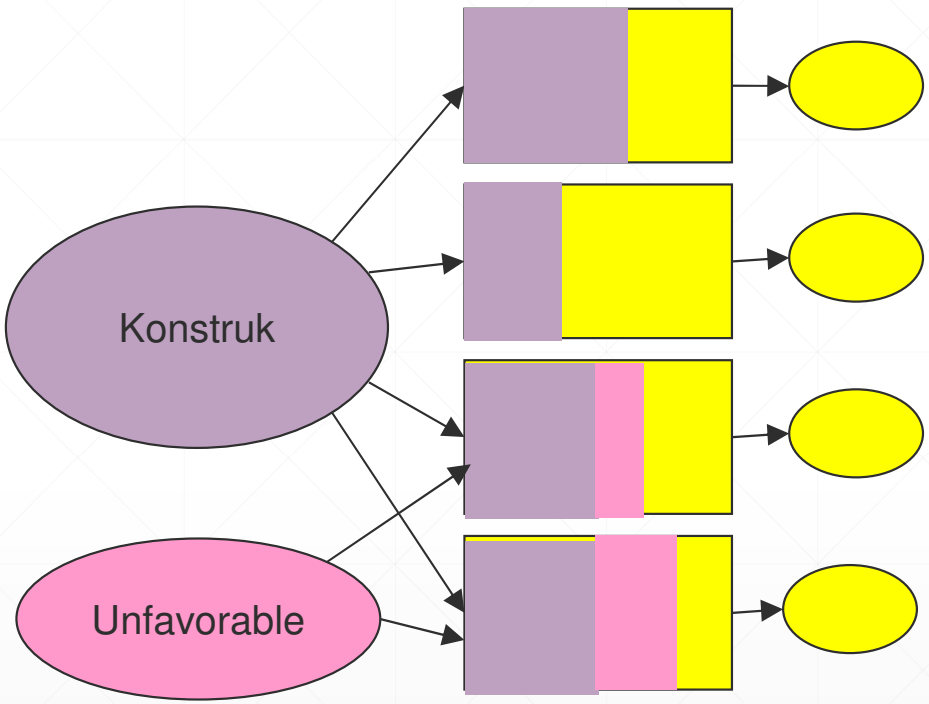
$t = \sigma^2$ trials
$r = \sigma^2$ examiner
$p = \sigma^2$ people

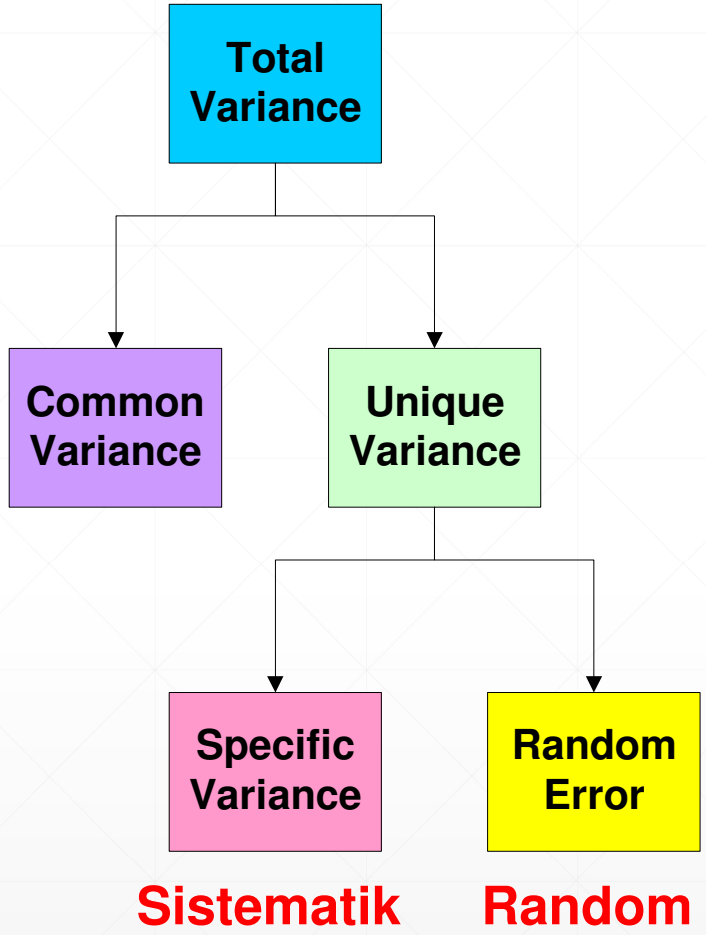
Kasus #3
Analisis Faktor



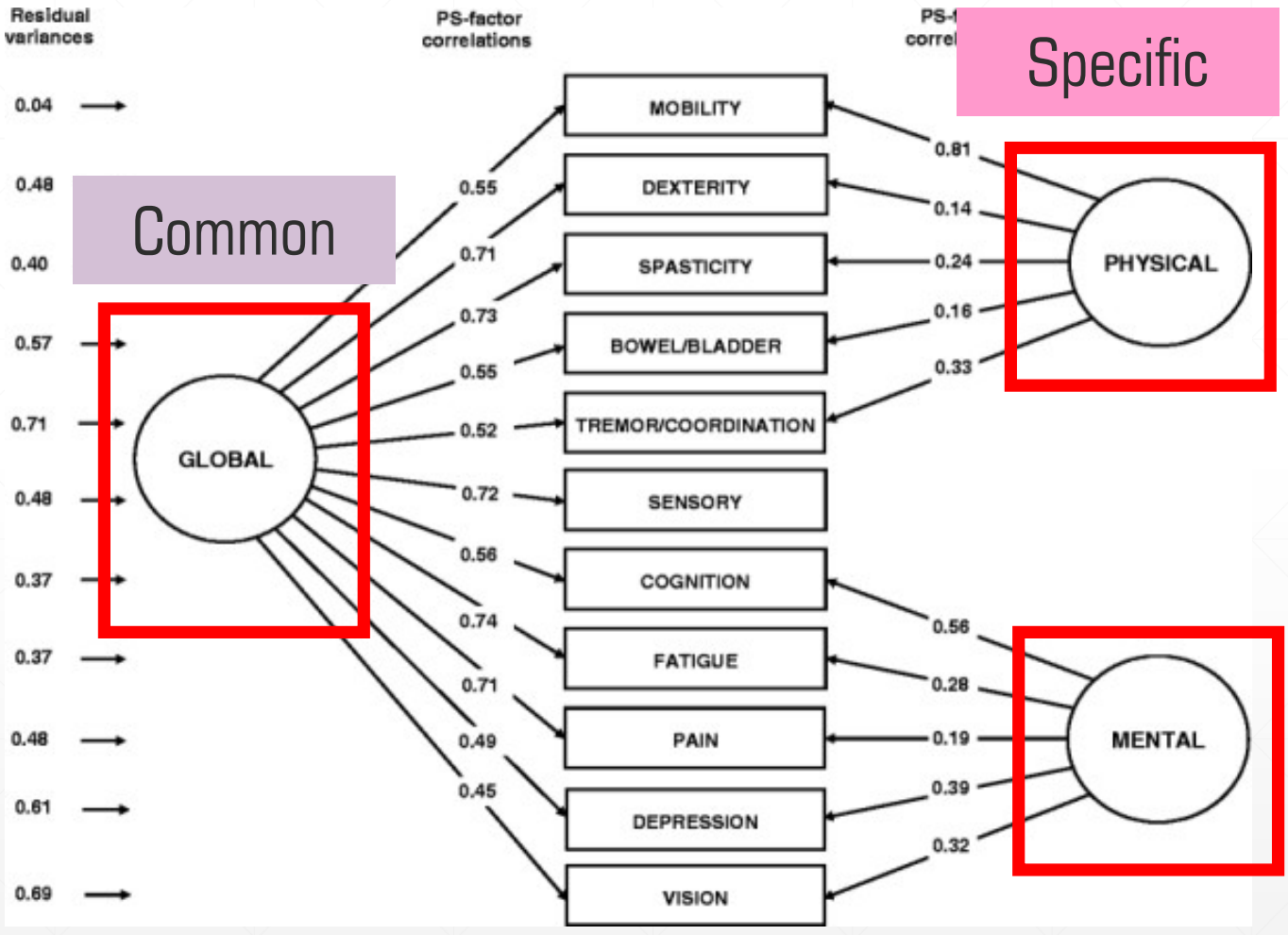


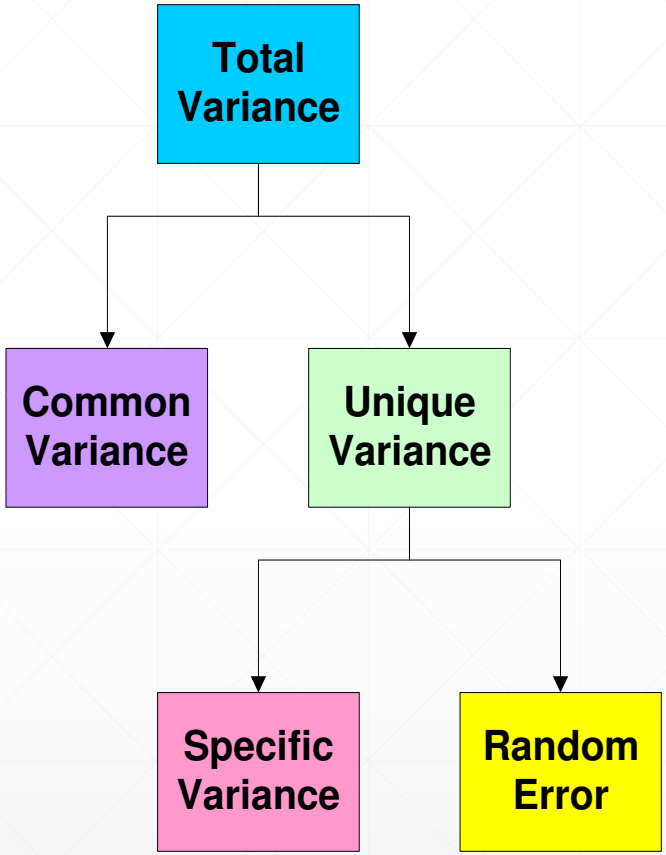
Gambar ini menunjukkan bahwa sebuah konstruk diukur oleh 4 indikator. Berdasarkan varians nya, skor dapat didekomposisi menjadi dua bagian, yaitu bagian **common variance** dan **unique variance**. Common variance terkait dengan konstruk yang diukur sedangkan yang unique terkait dengan eror pengukuran (varians acak yang tidak terjelaskan oleh model).





Gambar ini menunjukkan bahwa sebuah konstruk diukur oleh 4 indikator. Dua indikator memiliki kesamaan dalam hal arah butir (unfavorable) yang tidak terkait dengan konstruk ukur sehingga masuk ke dalam **unique variance**, oleh karena tidak bersifat random maka dinamakan dengan **specific variance**


Residual variances
PS-factor correlations
Specific
Common
0.04
0.48
0.40
0.57
0.71
0.48
0.37
0.37
0.48
0.61
0.69
GLOBAL
0.55
0.71
0.73
0.55
0.52
0.72
0.56
0.74
0.71
0.49
0.45
MOBILITY
DEXTERITY
SPASTICITY
BOWEL/BLADDER
TREMOR/COORDINATION
SENSORY
COGNITION
FATIGUE
PAIN
DEPRESSION
VISION
0.81
0.14
0.24
0.16
0.33
0.56
0.28
0.19
0.39
0.32
PHYSICAL
MENTAL


Total Variance
Common Variance
Unique Variance
Specific Variance
Random Error

Pendekatan Dasar (Konsistensi)

Cohen's Kappa (Konsistensi)

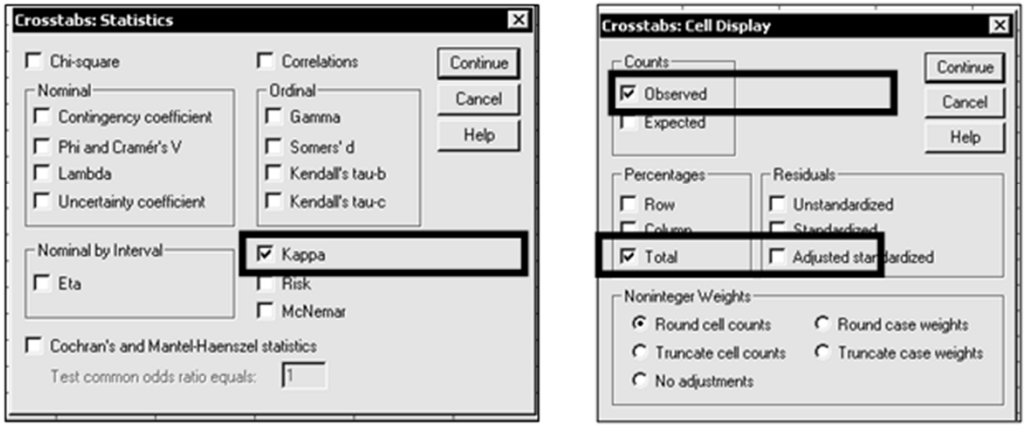
- Cohen (1960) mengembangkan koefisien untuk mengukur kesepakatan antar rater yang kemudian dikenal dengan koefisien kappa.
- **Penggunaan** Koefisien kappa tepat digunakan ketika (a) rater yang dipakai tidak banyak. Biasanya satu subjek dinilai oleh dua rater. (b) Skor hasil penilaiannya bersifat kategori. Biasanya juga hanya dua kategori yang dikode 0 atau 1.
- **Contoh Kasus** Dua orang psikolog menilai 10 orang klien apakah mereka perlu mendapatkan terapi ataukah tidak. Pada kasus ini terlihat jumlah raternya adalah dua dan kategori skor juga dua.
- **Contoh Ouput Analisis** Berikut ini contoh hasil analisis dengan menggunakan program SPSS. Hasil analisis menunjukkan kesepakatan antar rater sebesar 0.72. Tampak bahwa kedua psikolog memiliki indeks kesepakatan yang cukup tinggi.

Count

		Rater_B		
		Tidak Perlu Terapi	Perlu Terapi	Total
Rater_A	Tidak Perlu Terapi	5	1	6
	Perlu Terapi	1	8	9
Total		6	9	15

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Measure of Agreement	Kappa	.722	.183	2.797	.005
N of Valid Cases		15			

$\kappa < 0.00$ poor agreement
$0.00 < \kappa < 0.20$ slight
$0.21 < \kappa < 0.40$ fair
$0.41 < \kappa < 0.60$ moderate
$0.61 < \kappa < 0.80$ substantial, and
$0.81 < \kappa < 1.00$ almost perfect agreement.



Schmitz, et.al (2004) hendak menguji peranan informasi di media cetak dalam mempengaruhi persepsi penderita *attention deficit/hyperactivity disorder* (ADHD). Untuk memilih paragraf di dalam artikel yang akan digunakan dalam penelitian, peneliti memberikan 32 artikel media massa kepada 2 orang rater. Tugas rater adalah memilih paragraf mana saja dalam artikel tersebut yang relevan dengan pembahasan ADHD. Dari penelitian tersebut didapatkan nilai konsistensi yang tinggi antar kedua rater yang ditunjukkan dengan nilai kappa(κ)=0.92

		Rater_2		Jumlah
		Tipe A	Tipe B	
Rater_1	Tipe A	6	2	8
	Tipe B	0	12	12
Jumlah		6	14	20

		Rater_2		Jumlah
		Tipe A	Tipe B	
Rater_1	Tipe A	A	B	
	Tipe B	C	D	
Jumlah				

Nilai kappa didapatkan melalui transformasi tabel ke persamaan di bawah ini.

$$P_a = \frac{(A+D)}{(A+B+C+D)} = \frac{(6+12)}{(6+2+0+12)} = 0.900$$

dan $$P_c = \frac{(A+B)(A+C)(C+D)(B+D)}{(A+B+C+D)^2} = \frac{(6+2)+(6+0)+(0+12)+(2+12)}{(6+2+0+12)^2} = 0.540$$

$$\kappa = \frac{P_a - P_c}{1-P_c} = \frac{0.900-0.540}{1-0.540} = 0.783$$

Koefisien Korelasi intra Kelas (Konsistensi)

- Koefisien korelasi intra kelas (*intraclass correlation coefficients; ICC*) yang dikembangkan oleh Pearson (1901). Koefisien ini dikembangkan berdasarkan analisis varians namun pada kasus tertentu hasilnya memiliki kemiripan dengan koefisien alpha.
- **Penggunaan** Koefisien ICC tepat digunakan ketika (a) rater yang dipakai banyak dan (b) skor hasil penilaiannya bersifat kontinum.
- **Contoh Kasus** Enam orang psikolog menilai 10 orang anak apakah berapa banyak perilaku bullying yang mereka lakukan pada satu periode amatan. Contoh hasil pengamatan di tampilkan pada Tabel di bawah ini.

Subjek	Rater_1	Rater_2	Rater_k
1	1	2	1
2	4	4	3
3	3	3	3
n	5	5	6

Intraclass Correlation Coefficient

	Intraclass Correlation	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.811	.783	.837	35.287	299.0	2093	.000
Average Measures	.972	.967	.976	35.287	299.0	2093	.000

Nilai korelasi (ICC) didapatkan melalui rumus berikut, dengan menggunakan tabel ANOVA yang sudah di atas kita tidak kesulitan untuk mendapatkan nilai korelasi melalui rumus tersebut.

$$r = \frac{MS_{people} - MS_{residual}}{MS_{people} + (df_{people} \times MS_{residual})}$$

$$= \frac{0.473 - 0.011}{0.473 + (8 \times 0.11)} = 0.808$$

Hubungan antara ICC dengan alpha dapat diketahui melalui rumus berikut.

$$\alpha = \frac{k \times r}{1 + (k-1) \times r}$$

$$= \frac{9 \times 0.808}{1 + (9-1) \times 0.808}$$

$$= 0.974$$

Pendekatan Dasar (Ketepatan)

Indeks Ketepatan (Validitas)

- Aiken's Validity Indeks

$V = \sum s / [n(C-1)]$

$S = r - lo$

Lo = angka penilaian terendah (misalnya 1)

C = angka penilaian tertinggi (misalnya 4)

R = angka yang diberikan oleh penilai

- Lawshe's Content Validity Ratio

$$CVR = \frac{n_e - (N/2)}{N/2},$$

Angoff Method

Table 6-1 Hypothetical Data for Angoff Standard-Setting Method

Rater ID Number	*Item Number*													*Means (SD)*
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	
1	90 80	90 90	100 90	100 100	100 90	90 90	90 100	90 90	90 80	60 70	90 90	100 90	90 90	90.8 (10.38) 88.5 (8.01)
2	60 70	80 80	50 60	60 70	70 80	90 90	70 80	60 70	30 40	40 50	40 60	50 60	70 60	59.2 (17.06) 66.9 (13.78)
3	90 90	70 80	80 90	80 70	100 80	60 60	80 70	80 80	80 80	60 60	50 60	90 90	80 70	76.9 (13.77) 75.4 (11.27)
4	70 70	60 70	70 60	80 70	90 80	80 80	80 70	70 70	70 70	60 70	50 70	90 80	90 80	73.9 (12.61) 72.3 (5.99)
5	90 80	60 70	90 90	40 60	80 80	60 60	80 70	70 70	60 70	60 70	90 80	70 70	80 70	71.5 (15.19) 72.3 (8.32)
6	60 70	60 60	80 70	60 70	70 70	70 70	80 70	80 80	60 60	50 50	70 70	80 80	90 90	70.0 (11.55) 70.0 (10.00)
7	90 80	50 60	80 80	60 70	60 60	70 70	70 60	70 80	70 80	60 50	80 80	80 70	70 80	70.0 (10.80) 70.8 (10.38)
8	80 70	50 50	70 80	80 70	40 50	90 90	70 70	70 80	60 70	60 70	70 70	70 80	80 80	68.5 (13.45) 71.5 (11.44)
9	80 90	70 70	60 70	70 70	60 60	80 80	50 60	60 70	60 70	30 60	50 60	60 70	90 80	63.1 (15.48) 70.0 (9.13)
10	80 80	90 70	90 90	40 60	100 100	80 80	100 90	70 80	80 70	90 80	100 80	70 80	80 90	82.3 (16.41) 80.8 (10.38)
Means	79.0 78.0	68.0 70.0	77.0 78.0	67.0 71.0	77.0 75.0	77.0 77.0	77.0 74.0	72.0 77.0	66.0 69.0	57.0 63.0	69.0 72.0	76.0 77.0	82.0 79.0	72.6 (9.11) 73.9 (6.35)

Angoff Method (modified)

Table 6-2 Hypothetical Data and Example of Extended Angoff Standard-Setting Method

Item	*Rater ID Number*						*Means (SD)*
	1	*2*	*3*	*4*	*5*	*6*	
1	2 3	3 3	2 3	2 3	3 3	1 2	2.17 (.753) 2.83 (4.08)
2	1 2	2 2	1 2	2 2	2 3	1 2	1.50 (.548) 2.17 (.408)
3	2 3	2 3	2 3	2 3	2 3	2 2	2.00 (.000) 2.83 (.408)
4	3 3	3 3	2 3	2 3	3 3	2 3	2.50 (.548) 3.00 (.000)
5	1 2	1 2	2 2	1 2	2 2	1 1	1.33 (.516) 1.83 (.408)
6	2 3	3 3	3 3	2 3	3 3	2 2	2.50 (.548) 2.83 (.408)
7	3 3	2 3	2 3	2 3	3 3	2 3	2.33 (.516) 3.00 (.000)
8	2 3	3 3	3 3	2 3	3 3	2 3	2.50 (.548) 3.00 (.000)
Means (SD)	2.00 (.756) 2.75 (.518)	2.38 (.744) 2.75 (.463)	2.13 (.641) 2.75 (.463)	1.88 (.354) 2.75 (.463)	2.63 (.518) 2.88 (.354)	1.63 (.518) 2.25 (.707)	2.10 (.358) 2.69 (.221)

Angoff Method (modified)

Table 6-3 Hypothetical Data and Example of Yes/No Standard-Setting Method

	Rater ID Number						
Item	*1*	*2*	*3*	*4*	*5*	*6*	*Means (SD)*
1	1	0	0	1	0	1	0.50 (.224)
	1	1	0	0	0	1	0.50 (.224)
2	0	0	0	0	0	0	0.00 (.000)
	0	0	0	1	0	0	0.17 (.168)
3	1	1	0	1	1	1	0.83 (.168)
	1	1	0	1	1	1	0.83 (.168)
4	1	1	1	1	1	1	1.00 (.000)
	1	1	1	1	1	1	1.00 (.000)
5	0	0	0	0	0	0	0.00 (.000)
	0	0	0	0	0	0	0.00 (.000)
6	0	0	0	0	0	0	0.00 (.000)
	0	0	0	0	0	0	0.00 (.000)
7	1	1	1	1	1	1	1.00 (.000)
	1	1	1	1	1	1	1.00 (.000)
8	1	1	1	1	1	1	1.00 (.000)
	1	1	1	1	1	1	1.00 (.000)
9	1	1	1	1	1	1	1.00 (.000)
	1	1	1	1	1	1	1.00 (.000)
10	1	1	1	0	1	1	0.83 (.168)
	1	1	1	0	1	1	0.83 (.168)
11	0	0	0	0	0	0	0.00 (.000)
	0	0	0	0	0	0	0.00 (.000)
12	0	0	1	0	0	0	0.17 (.168)
	1	0	1	1	1	0	0.67 (.210)
Means	.58 (.149)	.50 (.151)	.50 (.151)	.50 (.151)	.50 (.151)	.58 (.149)	.53 (.041)
(SD)	.67 (.142)	.58 (.149)	.50 (.151)	.58 (.149)	.58 (.149)	.58 (.149)	.58 (.054)

Direct Consensus Method

Table 7-1 Hypothetical Data and Example of Direct Consensus Standard-Setting Method

	Rater ID Number									
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*		
Subtest Area (Number of Items)	*Participants' Recommended Number of Items Within a Subtest Area That the Just-Qualified Examinee Should Answer Correctly*								*Subtest Area Means* (SD)	*Percentage of Number of Items in Subarea*
A (9)	7	7	6	7	7	7	7	6	6.75 (0.46)	75.0
B (8)	6	6	7	6	5	7	7	7	6.38 (0.74)	79.8
C (8)	5	6	7	6	6	6	5	6	5.88 (0.64)	73.5
D (10)	7	8	8	7	7	8	7	7	7.38 (0.52)	73.8
E (11)	6	5	7	7	7	7	8	6	6.63 (0.92)	60.2
									Grand Mean *(SD)*	Percentage
Rater sums	31	32	35	33	32	35	34	32	33.00 (1.51)	71.7

Generalizability Theory

Tabel 2. Hasil Analis Varians Faset yang Terlibat di dalam Desain

Source	SS	df	MS	Random	Mixed	Corrected	%	SE
P	2.0025	2	1.0012	0.0011	0.0029	0.0029	0.5	0.0027
R	1.3136	2	0.6568	0.0014	0.0016	0.0016	0.3	0.0018
I	244.9580	89	2.7523	0.2301	0.2301	0.2275	40.4	0.0460
PR	0.9086	4	0.2272	0.0008	0.0025	0.0025	0.4	0.0015
PI	113.5531	178	0.6379	0.1619	0.1619	0.1619	28.7	0.0227
RI	34.9086	178	0.1961	0.0146	0.0146	0.0146	2.6	0.0079
PRI	54.2025	356	0.1523	0.1523	0.1523	0.1523	27.0	0.0114
Total	451.8469	809					100%	

Varians yang bersumber dari subjek (P) dan rater (R) cukup kecil yaitu 0,5% dan 0,3 %. Varians terbesar adalah butir tes (I) yang mendekati separuh dari varians dari semua faset (40.4%). Nilai interaksi antara subjek dan butir (PI) yang cukup besar (28.7%) menunjukkan bahwa subjek tertentu dapat mengatasi butir tertentu dibanding dengan subjek lainnya. Dengan kata lain, meskipun skor total yang dimiliki ketiga subjek adalah setara namun secara spesifik spesifikasi butir yang dapat diatasi oleh ketiga subjek adalah berbeda.

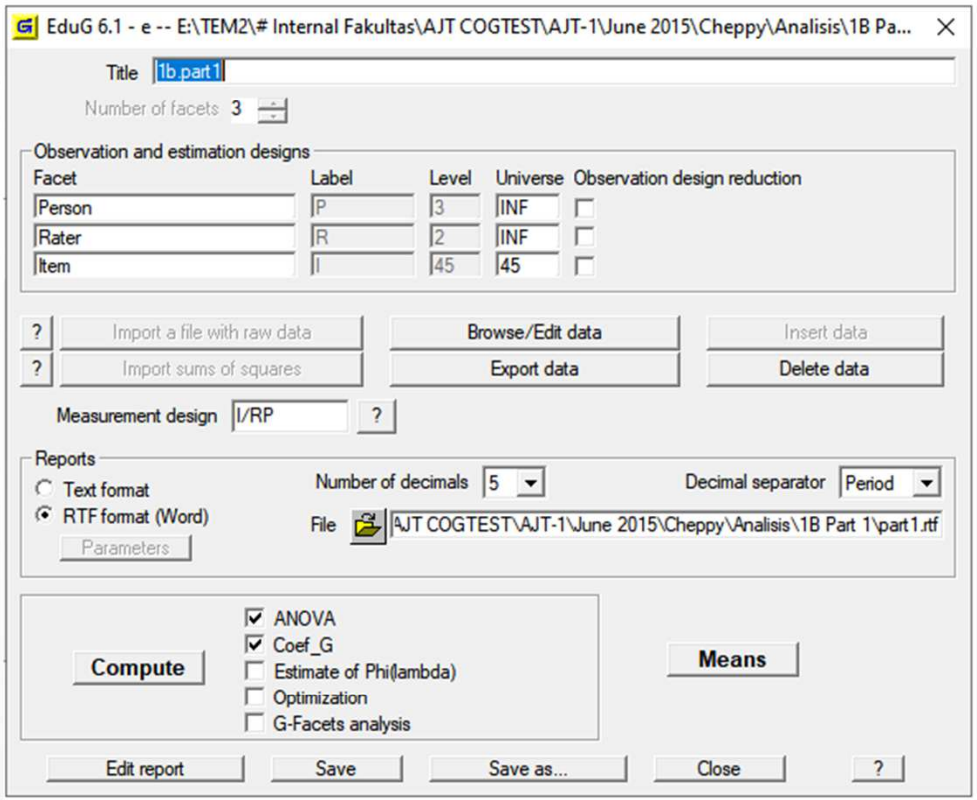

EduG 6.1 - e -- E:\TEM2\# Internal Fakultas\AJT COGTEST\AJT-1\June 2015\Cheppy\Analisis\1B Pa...
Title
1b.part1
Number of facets
3
Observation and estimation designs
Facet
Label
Level
Universe
Observation design reduction
Person
P
3
INF
Rater
R
2
INF
Item
I
45
45
?
Import a file with raw data
Browse/Edit data
Insert data
?
Import sums of squares
Export data
Delete data
Measurement design
I/RP
?
Reports
Text format
RTF format (Word)
Parameters
Number of decimals
5
Decimal separator
Period
File
AJT COGTEST\AJT-1\June 2015\Cheppy\Analisis\1B Part 1\part1.rtf
Compute
ANOVA
Coef_G
Estimate of Phi(lambda)
Optimization
G-Facets analysis
Means
Edit report
Save
Save as...
Close
?

Table 3.2 G Study Table for the Teaching Methods Example (Measurement Design M/SCI)

Source of variance	Differentiation variance	Source of variance	Relative error variance	% Relative	Absolute error variance	% Absolute
M	0.0273		—		—	
	—	C:M	0.0034	43.2	0.0034	29.6
	—	S:C:M	0.0012	15.5	0.0012	10.6
	—	I	—		0.0036	31.5
	—	Ml	0.0023	28.9	0.0023	19.8
	—	CI:M	0.0007	0.9	0.0007	6.2
	—	SI:C:M	0.0003	3.3	0.0003	2.3
Sum of variances	0.0273		0.0079	100%	0.0115	100%
Standard deviation	0.1653		Relative SE: 0.0889		Absolute SE: 0.1074	
Coef_G relative		0.78				
Coef_G absolute		0.70				

Many Facet Rasch

$$\ln\left[\frac{p(x_{1i}=1)}{p(x_{1i}=0)}\right]=\theta_1-\beta_i$$

examinees,
tasks

$$\ln\left[\frac{p_{nilvjk}}{p_{nilvjk-1}}\right]=\theta_n-\beta_i-\delta_l-\eta_v-\alpha_j-\tau_k,$$

examinees, tasks, interviewers, scoring criteria, and raters

MANY FACET Rasch

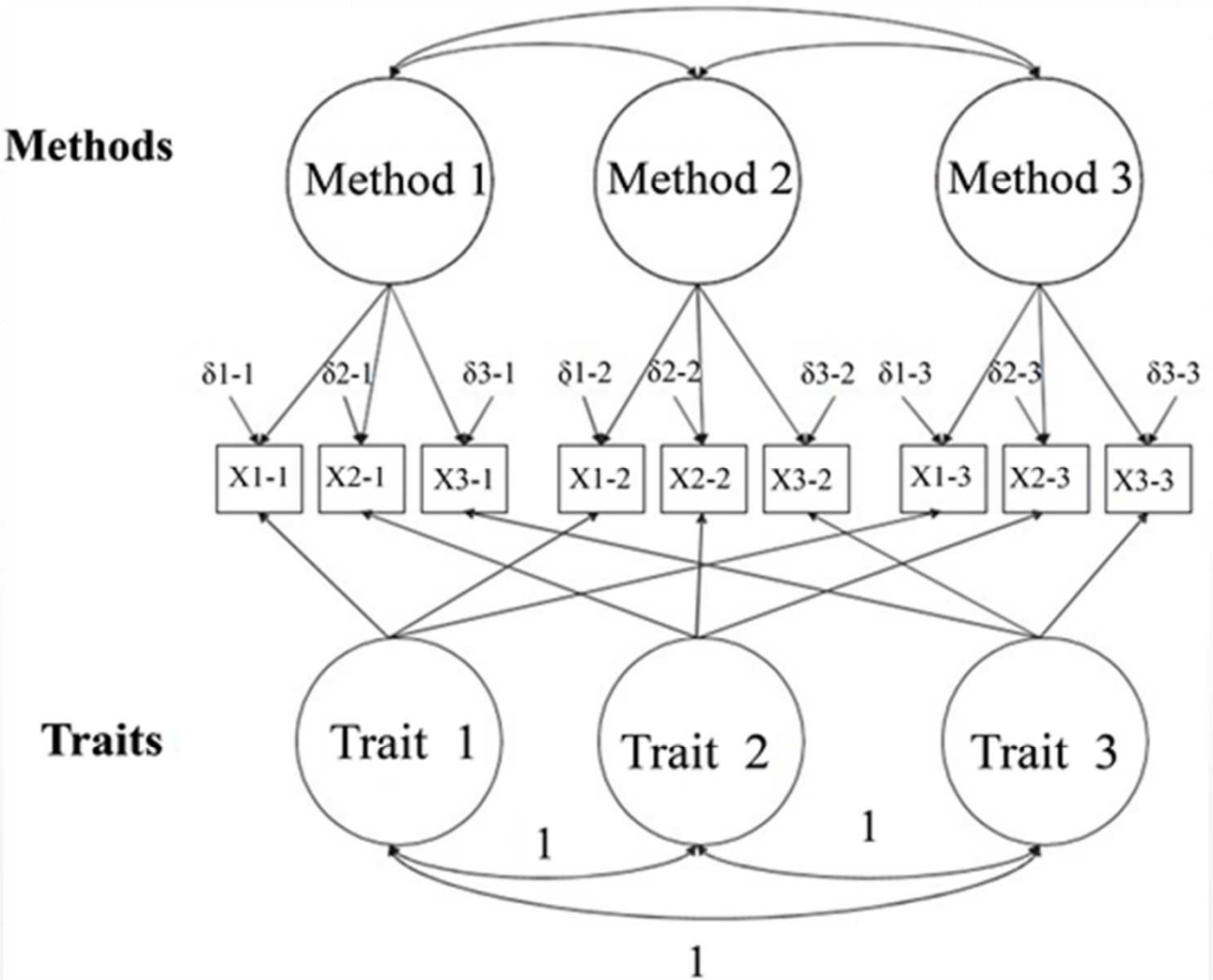
		Criterion		
Examinee	**Rater**	**GI**	**TF**	**LR**
001	01	5	5	4
001	14	3	3	3
002	07	4	5	3
002	10	4	4	4
…		.	.	.
…	.	.	.	.

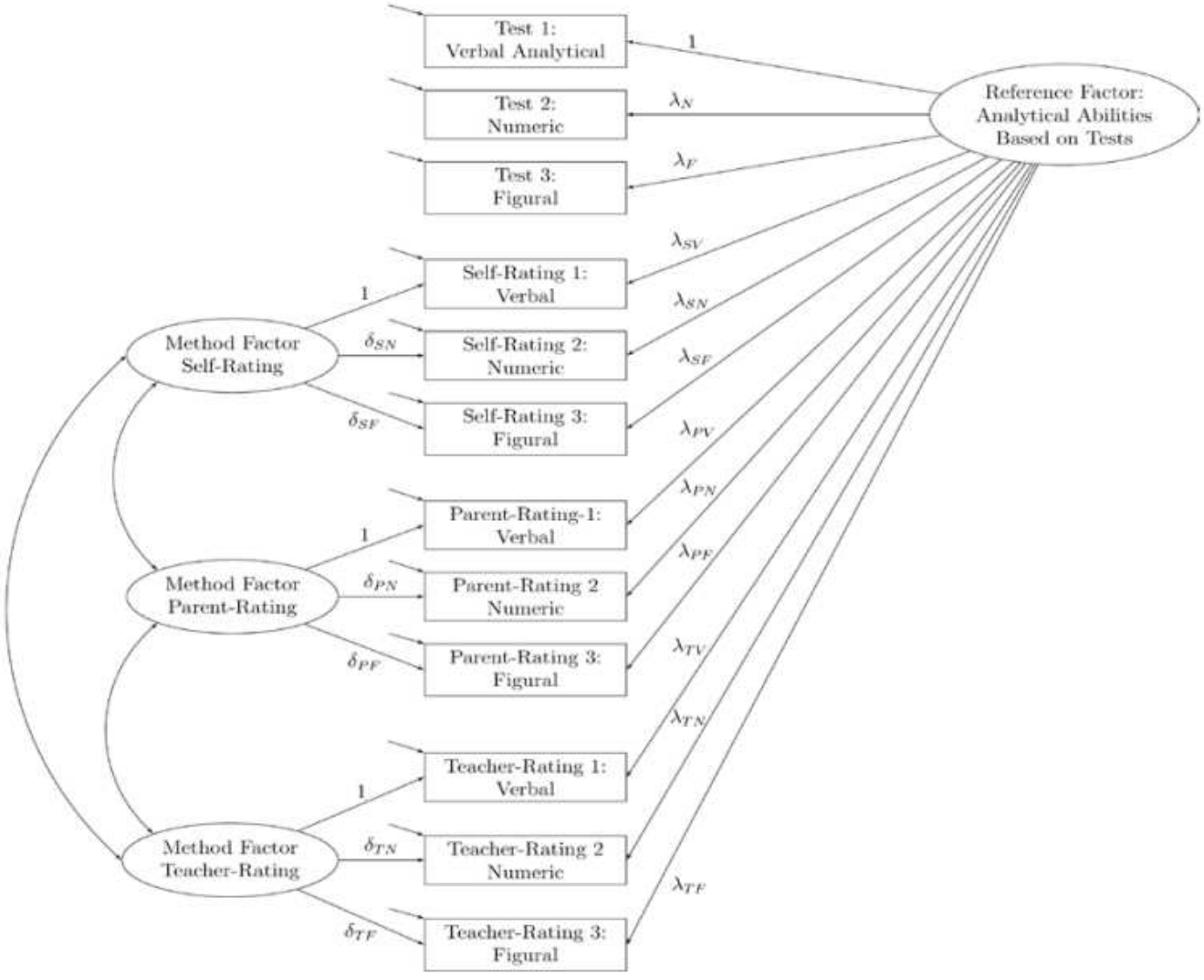
Table 3.1: Consensus and Consistency Indices of Interrater Reliability (Sample Data).

		Consensus Indices		Consistency Indices	
Rater Pair	***N***	**Exact Agreement**	**Cohen's Weighted Kappa**	**Pearson's *r***	**Kendall's Tau-b**
07 / 10	20	.70	.67	.83	.78
13 / 16	20	.60	.67	.84	.84
12 / 03	20	.55	.29	.49	.42

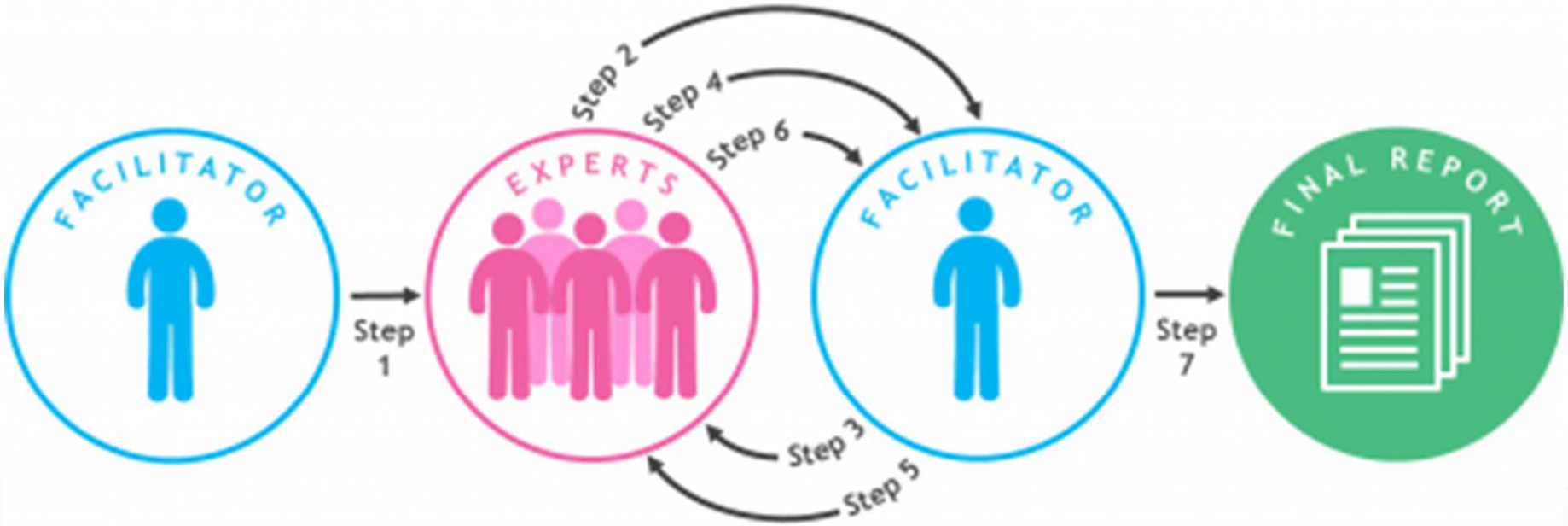
Logit	Examinees	Raters	Criteria	TDN Scale
	High	*Severe*	*Difficult*	(TDN 5)
	.			
7				
	.			
6	*.			
	*.			
	*.			
5	*			
	.			
	**			
4	***.			
	**.			-----
	**.			
3	**.			
	***			
	***	16		
2	*******.	13		
	***.	14		TDN 4
	***.	09 15		
1	*****	05		
	***.		LR	
	****.	04	TF	
0	***	06 08 11		-----
	*******	18		
	***.	17		
-1	***.	10 12	GI	
	***	02		
	****			
-2	**.	03		TDN 3
	**.	01 07		
	**.			
-3	**.			
	*.			
	**.			-----
-4	.			
	.			
	.			
-5	*			
	.			
	.			
-6	.			
	.			
	Low	*Lenient*	*Easy*	(below 3)

Multitrait Multirater via CFA


Methods
Method 1
Method 2
Method 3
δ1-1
δ2-1
δ3-1
δ1-2
δ2-2
δ3-2
δ1-3
δ2-3
δ3-3
X1-1
X2-1
X3-1
X1-2
X2-2
X3-2
X1-3
X2-3
X3-3
Traits
Trait 1
Trait 2
Trait 3
1
1
1


Test 1:
Verbal Analytical
Test 2:
Numeric
Test 3:
Figural
Reference Factor:
Analytical Abilities
Based on Tests
1
λ_N
λ_F
λ_{SV}
λ_{SN}
λ_{SF}
λ_{PV}
λ_{PN}
λ_{PF}
λ_{TV}
λ_{TN}
λ_{TF}
Method Factor
Self-Rating
Method Factor
Parent-Rating
Method Factor
Teacher-Rating
1
δ_{SN}
δ_{SF}
1
δ_{PN}
δ_{PF}
1
δ_{TN}
δ_{TF}
Self-Rating 1:
Verbal
Self-Rating 2:
Numeric
Self-Rating 3:
Figural
Parent-Rating-1:
Verbal
Parent-Rating 2
Numeric
Parent-Rating 3:
Figural
Teacher-Rating 1:
Verbal
Teacher-Rating 2
Numeric
Teacher-Rating 3:
Figural

Delphi Method



Facilitator seeks individual assessments from a pool of experts.

Experts respond to the request, receive feedback and revise their responses.

Facilitator compiles the responses and sends a revised set of questions to each expert. Several cycles of feedback may be needed.

Facilitator produces report on experts' responses, noting key outliers.

Final Report
Intermediate Summary
START: Assumptions & Key Questions
1. Delphi Round
Elaboration by experts
1. Expert Round
2. Delphi Round
Reexamining, commenting on and complementing the findings of the first round
2. Expert Round